

Н.А.Ниёзматова, Д.Т.Мухамедиева

*Центр разработки программных продуктов и аппаратно-программных комплексов при Ташкентском университете информационных технологий,
г.Ташкент*

УДК 519.71(575.1)

РЕШЕНИЕ ЗАДАЧИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ НА ОСНОВЕ ПРИМЕНЕНИЯ ИСКУССТВЕННЫХ ИММУННЫХ СИСТЕМ

Аннотация. Решение задач интеллектуального анализа данных и принятия решений в слабоформализуемых системах только с помощью детерминированных и вероятностных методов является недостаточным. Для этого необходимо широко применять методы гибридных интеллектуальных систем и, особенно, методы «мягких» вычислений. В работе рассматривается решение задачи интеллектуального анализа данных с использованием теории нечетких множеств и иммунных алгоритмов.

Ключевые слова: Интеллектуальный анализ данных, нечеткое множество, иммунные алгоритмы, классификация, клонирование, мутация.

1. Введение

В настоящее время, в связи с достаточно быстрым развитием информационных систем, от лица принимающего решение требуется не только профессиональное знание технологического процесса, основ управления, но и опыт работы в этих информационных системах, умение принимать решение.

Выбор конкретной модели из систем интеллектуального анализа данных и требуемая информационная емкость определяются, прежде всего, потребностью обеспечения определенной степени «наблюдаемости» конкретного объекта. При этом все вопросы, касающиеся оперативного решения более сложных задач, выходящих за рамки традиционной обработки информации, остаются вне «поля зрения» этих систем.

К числу задач, стоящих перед лицом, принимающим решения (ЛПР), относятся [1-8]:

- анализ проблемной ситуации,
- идентификация возникшего отклонения от нормального (штатного) режима функционирования объекта,

- поиск возможных корректирующих решений по воздействию на объект,
- прогнозирование ситуаций,
- оценка последствий принимаемых решений,
- выдача команд на обработку необходимых управляющих воздействий.

Различные стратегии интеллектуализации мониторинга направлены на реализацию интеллектуальной информационной поддержки ЛПР, использующего средства мониторинга. Такую поддержку можно реализовать путем построения нечетких лингвистических баз данных/знаний вместе с подсистемами нечеткого вывода, причем информация для принятия решений может выводиться на автоматизированное рабочее место ЛПР [9-15].

2. Постановка задачи

Нечеткие логические уравнения задачи классификации (типа если <вход>, то <выход>) вместе с функциями принадлежности нечетких термов позволяют принимать решение с использованием следующего алгоритма:

1 Фиксируются значения параметров состояния объекта [16-17]:

$$X = [x_1, \dots, x_n].$$

2 Определяются значения функций принадлежности $\mu^j(x_i^*)$, при фиксированных значениях параметров x_i^* , $i = \overline{1, n}$.

3 Используя логические уравнения, вычисляются значения функций принадлежности $\mu^{r_j}(x_1^*, x_2^*, \dots, x_n^*)$ при векторе состояния $X^* = (x_1^*, x_2^*, \dots, x_n^*)$.

4 Определяется решение r_j^* , для которого:

$$\mu^{r_j^*}(x_1^*, x_2^*, \dots, x_n^*) = \max_{j=1, n} [\mu^{r_j}(x_1^*, x_2^*, \dots, x_n^*)].$$

Суть обучения состоит в подборе таких параметров функций принадлежности, которые минимизируют различие между результатами нечеткой аппроксимации и реальным поведением объекта.

3. Этапы процесса обучения

Процесс обучения состоит из следующих этапов:

1 Сформировать обучающую матрицу $A = [X_1, \dots, X_m]^T$ размерности $m \times n$.

2 Вычислить максимальное сингулярное число s , а также левый и правый сингулярные векторы L и R обучающей матрицы по следующей итеративной (эволюционной) схеме:

$$L_{(0)} = [1, \dots, 1]^T,$$

$$R^T = L_{(k-1)}^T A, \quad R_{(k)} = R / |R|, \quad \text{где } |R| = \sqrt{r_1^2 + \dots + r_n^2},$$

$$L = AR_{(k)}, \quad L_{(k)} = L / |L|, \quad \text{где } |L| = \sqrt{l_1^2 + \dots + l_n^2},$$

$$s_{(k)} = L_{(k)}^T AR_{(k)}, \quad k = 1, 2, \dots,$$

до выполнения условия $|s_{(k)} - s_{(k-1)}| < \varepsilon$,

$$s = s_{(k)}, \quad L = L_{(k)}, \quad R = R_{(k)}.$$

3 Хранить сингулярное число s .

4 Хранить правый сингулярный вектор R (как «антитело-пробу»).

5 Для всякого $i=1, \dots, m$ хранить компоненту l_i левого сингулярного вектора L и класс c_i соответствующий обучающему образу X_i .

6 Для всякого n -мерного образа Z вычислить его энергию связи с R :

$$w(z) = Z^T R / s.$$

7 Выбрать l_i , которая имеет минимальное расстояние d с w :

$$d = \min_i |w - l_i|, \quad i=1, \dots, m,$$

и считать класс c_i искомым классом образа Z .

В общем виде один шаг работы иммунного алгоритма можно представить следующим образом [6]:

$$\begin{aligned} \forall i \in \{1, \dots, m\} : x'_i &= \text{mut}(\text{clon}(x_1, \dots, x_n)) \\ (x''_1, \dots, x''_k) &= \text{aging}(x_1, \dots, x_n, x'_1, \dots, x'_m) \\ (y_1, \dots, y_n) &= \text{sel}(x''_1, \dots, x''_k) \end{aligned} \quad (1)$$

где $(x_1, \dots, x_n) \in X^n$ – текущая популяция антител; (x'_1, \dots, x'_m) – популяция антител, возникающая в результате клонирования и мутации; (x''_1, \dots, x''_k) – антитела, которые удаляются из популяции; (y_1, \dots, y_n) – антитела, добавляемые в текущую популяцию.

В развитие работы планируется исследование эффективности метода в условиях высокой размерности вектора варьируемых параметров, а также более широкое исследование влияния свободных параметров метода на его эффективность. Планируется также разработка и исследование модификаций метода, ориентированных на параллельные вычислительные системы различной архитектуры (кластеры, системы с общей памятью, графические процессорные устройства). Наконец, имеется в виду сравнение эффективности и интеграция различных поведенческих методов [8, 18-20].

4. Применение иммунного алгоритма для обучения нечеткой модели

В практике разработки интеллектуальных систем имеется тенденция использования гибридных моделей для решения различных практических задач. Так, применительно к обучению нечетких моделей некоторые авторы, наряду с традиционными методами оптимизации, используют иммунный алгоритм [8]. Экспериментальные исследования, проведенные авторами, показывают, что использование данного подхода позволяет добиться лучших результатов по повышению скорости и точности решаемых задач. Таким образом, применительно к проблеме формирования баз знаний экспертных систем актуальной задачей является разработка эффективного алгоритма обучения нечеткой модели на основе интеллектуальных методов и эффективных эвристических алгоритмов.

Пусть задана выборка нечетких экспериментальных данных (X_r, y_r) , $r = \overline{1, M}$; здесь $X_r = (x_{r1}, x_{r2}, \dots, x_{rm})$ - входной n -мерный вектор и $y_r = (y_1, y_2, \dots, y_M)$ - соответствующий ему выходной вектор.

В общем виде требуется построить модель, основанную на нечетких правилах вывода:

$$\bigcup_{p=1}^{k_j} \left(\bigcap_{i=1}^n x_i = a_{i,jp} - \text{с весом } w_{jp} \right) \rightarrow \\ \rightarrow y_j = b_{m0} + b_{m1}x_1^j + \dots + b_{mn}x_n^j.$$

В процессе построения модели нужно найти такие значения коэффициентов правил

$$B = (b_{ij}), i = \overline{1, m}, j = \overline{0, n}$$

при которых достигается минимум следующего выражения:

$$\sum_{r=1}^M (y_r - y_r^f) \rightarrow \min, \quad (2)$$

где y_r^f - результат нечетких правил вывода с параметром B в r -й строке выборки (X_r) .

Алгоритм настройки параметров функций принадлежности $B = (b_1, b_2, \dots, b_q)$ и $C = (c_1, c_2, \dots, c_q)$ - и веса правил $W = (w_1, w_2, \dots, w_N)$ состоит из следующих этапов.

1 Формирование исходной популяции. *Оператор клонирования* генерирует новое поколение копий антител в будущей популяции. Известны следующие основные операторы клонирования: а) статический оператор клонирования, который просто копирует каждую В-клетку, производя переходную популяцию; б) пропорциональный оператор клонирования, который клонирует В-клетки пропорционально их антигенной схожести; в) оператор вероятностного клонирования, в соответствии с которым В-клетки выбираются из текущего поколения в зависимости от вероятности клональной селекции [7].

Для реализации ИА следует задать способ кодирования нечетких моделей. Сведем неизвестные параметры W, B, C в один вектор:

$$S = (W, B, C) = (w_1, w_2, \dots, w_N, b_{11}, c_{11}, \dots, b_{l_1}, c_{l_1}, b_{n1}, c_{n1}, \dots, b_{nl_1}, c_{nl_1})$$

где N - общее число строк в нечеткой базе знаний;

l_i - количество термов-оценок входной переменной λ_i ,

$$l_1 + l_2 + \dots + l_n = q, \quad i = \overline{1, n};$$

q - общее число термов;

2 *Оператор мутации* действует в зависимости от имеющейся популяции клонов, применяя к каждому антителу определенное количество одиночных мутаций, осуществляемых случайным образом.

Каждый элемент вектора S может подвергнуться операции мутации с вероятностью p_m . Обозначим мутацию элемента s через $Mu(s)$:

$$Mu(w_j) = RANDOM\left(\left[\underline{w}, \overline{w}\right]\right), j = \overline{1, N}$$

$$Mu(b_{ip}) = RANDOM\left(\left[\underline{x_i}, \overline{x_i}\right]\right),$$

$$Mu(c_{ip}) = RANDOM\left(\left[\underline{c_i}, \overline{c_i}\right]\right),$$

где $\underline{w}, (\overline{w})$ - нижняя (верхняя) граница интервала возможных значений весов правил, $[\underline{w}, \overline{w}] \subset [0, 1]$;

$[\underline{\tilde{n}_i}, (\overline{c_i})]$ - интервал возможных значений коэффициента концентрации-растяжения функций принадлежности термов-оценок входной переменной x_i , $[\underline{\tilde{n}_i}, (\overline{c_i})] \subset (0, +\infty]$, $i = \overline{1, n}$;

$RANDOM\left(\left[\underline{\xi}, \overline{\xi}\right]\right)$ обозначает операцию нахождения равномерно распределенного на интервале $[\underline{\xi}, \overline{\xi}]$ случайного числа.

3 Оператор старения устраняет старые особи. Статический оператор старения использует возрастной параметр для максимального количества поколений антител, которым разрешено оставаться в популяции. Когда антитело старше, оно удаляется из системы, даже если оно может оказаться вполне пригодным на последующих итерациях.

При клональной экспансии клонированное антитело наследует возраст его родителя. После этапа мутации только те антитела, которые получили высшее значение аффинности, получают возраст, равный 0. Элитный вариант этого оператора получается путем взятия наилучших антител популяции в поколение с возрастом, равным 0.

Элитный вариант этого оператора получается взятием лучшего антитела из популяции в поколение.

4 Оператор селекции заменяет наихудшие антитела в популяции новыми случайными антителами.

На основе описанного ИА разработано программное обеспечение для решения практических задач многокритериальной оптимизации и получены результаты оптимизационной задачи [5,9].

Для проведения сравнительного анализа взяты известные модельные задачи, расположенные по электронному адресу: <http://www.ics.uci.edu/~mllearn/databases/>. К ним относятся следующие задачи: задача ирис (Iris Data Set), задача идентификации стекла (Glass Identification Data Set), задача диагностирования диабета (Pima Indians Diabetes), экологическая задача (Ecoli Data Set), задача Хабермана (Haberman's Survival Data Set), задача определения сорта вина (Wine Data Set), задача определения состояния печени (Liver).

Ниже в таблице 1 приводятся параметры перечисленных задач.

Таблица 1

Параметры модельных задач

Наименование задачи	Количество классов	Количество признаков	Количество объектов
Стекло (Glass)	7	9	214
Хаберман (Haberman)	2	4	306
Ирис (Iris)	3	4	150
Экология (Ecoli)	8	7	336
Вино (Wine)	3	13	178
Печень (Liver)	2	6	345

В таблице 2 для сравнения приведены результаты решения некоторых модельных задач на основе различных известных и предлагаемого алгоритмов.

Таблица 2

Результаты работы предлагаемого и существующих алгоритмов

Задача	Предлагаемый алгоритм	GBC	SGF	SVM	1NN	KNN	Conventional RBF network
Стекло (Glass)	87.2	84.27	75.74	71.50	72.01	72.01	69.16
Ирис (Iris)	98.5	98.00	97.33	97.33	96.00	95.33	95.33
Вино (Wine)	100	100	99.44	99.44	95.52	96.07	98.89

В таблице 3 приведены лучшие результаты работы рассматриваемых алгоритмов. Наиболее лучшие результаты для конкретной задачи выделены отдельно.

Кроме того, в таблице 3 сравниваются различные результаты работы предлагаемого нами алгоритма, алгоритмов GBC (Gravitation Based Classification) и SVM (Support Vector Machines).

Таблица 3

Результаты работы предлагаемого и существующего алгоритмов

	Предлагаемый метод			GBC			SVM		
	Очень плохо	Очень хорошо	Плохо	Очень плохо	Очень хорошо	Плохо	Очень плохо	Очень хорошо	Плохо
Хаберман (Haberman)	82.3	91.6	87.2	75.2	86.7	81.5	72.3	82.1	78.8
Печень (Liver)	77.5	85.1	82.4	63.5	72.8	67.1	60.4	68.3	65.5
Экология (Ecoli)	89.7	94.3	92.7	89.7	98.5	95.5	89.4	94.4	92.3

5. Заключение

Таким образом, показана целесообразность объединения метода нечеткого вывода и иммунных алгоритмов в задачах с параметрами, содержащими неопределенности различного типа, а также в задачах, для

которых характерны интуитивные решения. Предложенный метод позволяет существенно улучшить качество решения многокритериальных задач оптимизации с нечетко заданными параметрами и критериями. В дальнейшем планируется изучение различных гибридных методов применительно к оптимизационным задачам, а также методов автоматического формирования базы нечетких правил. Это позволит существенно повысить эффективность решений оптимизационных задач рассматриваемого класса.

СПИСОК ЛИТЕРАТУРЫ

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. 2-е изд. Уч. пособие. СПб.: БХВ – Петербург, 2007. – 384 с.: ил. + CD-ROM.
2. Aliev R.A., Aliev R. Theory of Intelligent Systems and Applications / -Baku: Chashyogly, 2001. -720 p.
3. Zaychenko Yu. The Fuzzy Group Method of Data Handling and Its Application for Economical Processes forecasting // Scientific Inquiry. -Vol. 7. -№1, June, 2006. -pp. 83-98.
4. Кондрашина Е.Ю., Литвинцева Л.В., Поспелов Д.А. Представление знаний о времени и пространстве в интеллектуальных системах. Под ред. Д.А. Поспелова. Серия «Проблемы искусственного интеллекта», -Вып. 6. – М.: Наука. Физматгиз, 1989. -328 С.
5. Kennedy J., Eberhart R. C. “Particle swarm optimization” // In Proceedings of IEEE International Conference on Neural Networks. 1995. -pp. 1942–1948.
6. Zadeh L.A. What is Soft Computing? Soft Computing 1 (1997).
7. Дасгупта Д. Искусственные иммунные системы и их применение.- Физматлит.- 2006.-344 с.
8. Dasgupta D., Artificial Immune Systems and Their Applications, Springer-Verlag, 1998.

9. Jie J., Han Ch., Zeng J. An Extended Mind Evolutionary Computation Model for Optimizations // Applied Mathematics and Computation, 2007, No. 185(2), pp. 1038 – 1049
10. Бекмуратов Т.Ф., Дадабаева Р. А., Мухамедиева Д.Т. Принятие решений в нечеткой среде // Научный журнал «Проблемы информатики» – Новосибирск, Институт вычислительной математики и математической геофизики СО РАН, 2010, вып.1, с. 52-61.
11. Мухамедиева Д.Т. Информационно-диалоговая система «Определение нечетких параметров имитационной модели» Свидетельство № DGU 00879 об официальной регистрации программы для ЭВМ в Государственном Патентном ведомстве. - Ташкент, 2005.
12. Мухамедиева Д. Т., Солиева Б. Т. Алгоритм решения задачи линейного программирования с использованием теории нечётких множеств // М.: Актуальные проблемы современной науки, 2010. – С. 137-139.
13. Бекмуратов Т.Ф., Мухамедиева Д.Т., Бобомурадов О. Нечеткий алгоритм принятия слабоструктурированных решений // Актуальные проблемы современной науки. –Москва. 2010. №5. С. 124-128.
14. Мухамедиева Д.Т., Примова Х.А. Модели задач параметрической оптимизации слабоформализуемых процессов // Актуальные проблемы современной науки. –Москва. 2010. №5. С. 134-137.
15. Мухамедиева Д.Т., Мингликулов З.Б. Решение задачи оптимального исследования рынка с применением нейронных сетей // Актуальные проблемы современной науки. –Москва. 2010. №5. С. 131-134.
16. Мухамедиева Д.Т., Агзамходжаева М.Р. Решение задачи оптимального использования торговых агентов с помощью генетического алгоритма // Актуальные проблемы современной науки. –Москва. 2010. №5. С. 128-131.
17. Мухамедиева Д.Т., Солиева Б.Т. Алгоритм решения задачи линейного программирования с использованием теории нечетких множеств // Актуальные проблемы современной науки. –Москва. 2010. №5. С. 137-140.

18. Бекмуратов Т.Ф., Мухамедиева Д.Т., Бобомурадов О.Ж.. Нечеткая модель прогнозирования урожайности //Научный журнал «Проблемы информатики» – Новосибирск, Институт вычислительной математики и математической геофизики СО РАН, 2010, вып.3. С. 11-23.
19. Мухамедиева Д.Т. Алгоритм кластеризации правил систем нечеткого вывода // Естественные и технические науки. №2. -Москва. 2013. -С.248-252.